NAACL-HLT 2021

Multimodal Artificial Intelligence (MAI)

Proceedings of the Third Workshop

June 6, 2021

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-25-1

The NAACL 2021 Workshop on Multimodal Artificial Intelligence (MAI-Workshop) offers a unique opportunity for interdisciplinary researchers to study and model interactions between (but not limited to) modalities of language, vision, and acoustic. Advances in multimodal learning allows the field of NLP to take the leap towards better generalization to real-world (as opposed to limitation to textual applications), and better downstream performance in Conversational AI, Virtual Reality, Robotics, HCI, Healthcare, and Education. We invite researchers from NLP, Computer Vision, Speech Processing, Robotics, HCI, and Affective Computing to submit their papers.

• Neural Modeling of Multimodal Language
• Multimodal Dialogue Modeling and Generation
• Multimodal Sentiment Analysis and Emotion Recognition
• Language, Vision and Speech
• Multimodal Artificial Social Intelligence Modeling
• Multimodal Commonsense Reasoning
• Multimodal RL and Control (Human-robot communication and multimodal language for robots)
• Multimodal Healthcare
• Multimodal Educational Systems
• Multimodal Affective Computing
• Multimodal Fusion and Alignment
• Multimodal Representation Learning
• Multimodal Sequential Modeling
• Multimodal Co-learning and Transfer Learning
• Multimodal Active Learning
• Multimodal and Multimedia Resources
• Creative Applications of Multimodal Learning in E-commerce, Art, and other Impactful Areas.

Amir Zadeh – Language Technologies Institute, Carnegie Mellon University
Louis-Philippe Morency – Language Technologies Institute, Carnegie Mellon University
Paul Pu Liang – Machine Learning Department, Carnegie Mellon University
Candace Ross – Massachusetts Institute of Technology
Ruslan Salakhutdinov – Carnegie Mellon University
Soujanya Poria – Singapore University of Technology and Design
Erik Cambria – Nanyang Technological University
Kelly Shi – Carnegie Mellon University

Table of Contents

Conference Program

Multimodal Weighted Fusion of Transformers for Movie Genre Classification
Isaac Rodríguez Bribiesca, Adrián Pastor López Monroy and Manuel Montes-y-Gómez

On Randomized Classification Layers and Their Implications in Natural Language Generation
Gal-Lev Shalev, Gabi Shalev and Joseph Keshet

COIN: Conversational Interactive Networks for Emotion Recognition in Conversation
Haidong Zhang and Yekun Chai

A First Look: Towards Explainable TextVQA Models via Visual and Textual Explanations
Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian and Mingwei Shen

Multi Task Learning based Framework for Multimodal Classification
Danting Zeng

Validity-Based Sampling and Smoothing Methods for Multiple Reference Image Captioning
Shunta Nagasawa, Yotaro Watanabe and Hitoshi Iyatomi

Modality-specific Distillation
Woojeong Jin, Maziar Sanjabi, Shaoliang Nie, Liang Tan, Xiang Ren and Hamed Firooz

Cold Start Problem For Automated Live Video Comments
Hao Wu, François Pitie and Gareth Jones

¡Qué maravilla! Multimodal Sarcasm Detection in Spanish: a Dataset and a Baseline
Khalid Alnajjar and Mika Hämäläinen

A Package for Learning on Tabular and Text Data with Transformers
Ken Gu and Akshay Budhkar

Semantic Aligned Multi-modal Transformer for Vision-LanguageUnderstanding: A Preliminary Study on Visual QA
Han Ding, Li Erran Li, Zhiting Hu, Yi Xu, Dilek Hakkani-Tur, Zheng Du and Belinda Zeng

GraghVQA: Language-Guided Graph Neural Networks for Graph-based Visual Question Answering
Weixin Liang, Yanhao Jiang and Zixuan Liu

No Day Set (continued)

Learning to Select Question-Relevant Relations for Visual Question Answering
Jaewoong Lee, Heejoon Lee, Hwanhee Lee and Kyomin Jung